.

‘Evaluating ChatGPT Responses on Atrial Fibrillation for Patient Education’ Çalışma Değerlendirmesi

Dr. Hidayet Ozan Arabacı

‘Evaluating ChatGPT Responses on Atrial Fibrillation for Patient Education’ Çalışma Değerlendirmesi

Dr. Hidayet Ozan Arabacı
İstanbul-Üniversitesi Cerrahpaşa Kardiyoloji Enstitüsü

1) Çalışmanın Adı: Evaluating ChatGPT Responses on Atrial Fibrillation for Patient Education1
2)Çalışmanın Yayınlandığı Dergi: Cureus- Springer Nature

3)Çalışmanın Yayınlandığı Tarih: 4 Haziran 2024

4)Çalışmanın Sponsoru: Çalışmanın herhangi bir kişi ya da kuruluş sponsoru bulunmamaktadır.

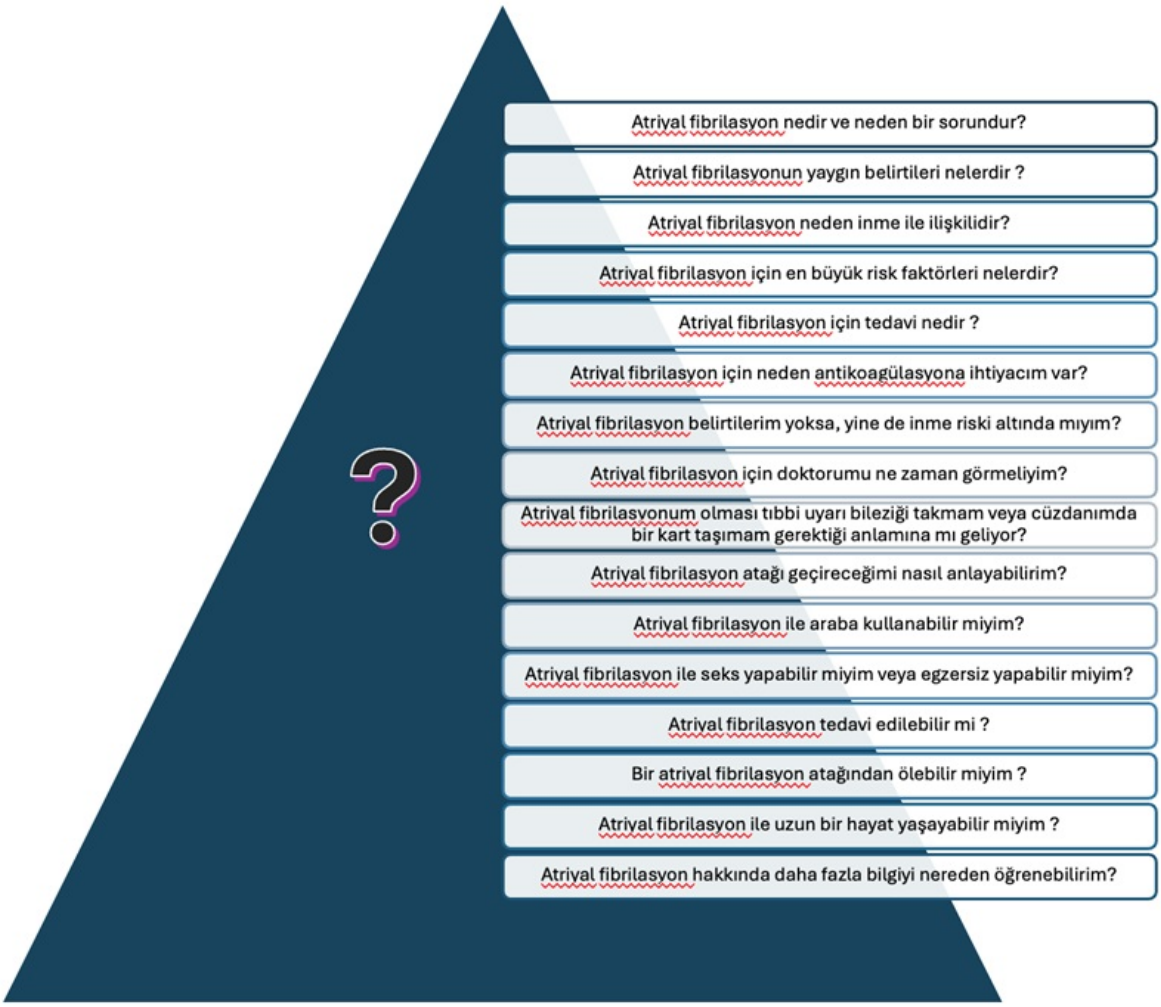
5)Çalışmanın Amacı:
OpenAI tarafından geliştirilen bir yapay zeka (AI) sohbet robotu olan ChatGPT, Kasım 2022'de piyasaya sürüldü ve sonrasında yaygın bir ilgiyle artarak kullanıma girdi. Yeni geliştirilen bir teknolojiden internet aramaları için yaygın olarak kullanılan bir araca dönüştü. AI tabanlı sistemlerin sağlık alanında hastalar ve profesiyonel sağlık çalışanlarına gerek hasta eğitimi gerek tanı ve tedavi algoritmaları açısından sağladığı fayda giderek artmaktadır ve bu sistemlerin kullanımını daha da yaygınlaştırmaktadır. Özellikle AI tabanlı sohbet robotu olan sistemlerin tıbbi hastalıklar, bulguları, erken tedavi imkanı yaratmaları ve farkındalığı arttırması açısından önemli günlük kullanımda önemli bir hale gelmektedir. Bu sebeble ChatGPT'nin AF ile ilgili yanıtlarının kalitesini ve doğruluğunu değerlendirmek çok önemlidir. Bu çalışmada, ChatGPT'nin; hastaların AF ile ilgili sorularına verdiği yanıtları eleştirel bir bakış açısıyla değerlendirmeyi ve bu yanıtların doğruluğuna, netliğine ve hasta eğitimi için uygunluğuna odaklanmayı amaçlamaktadır. Elde edilen veriler, ChatGPT gibi yapay zeka araçlarının hasta eğitimindeki potansiyel uygulamalarını ve sınırlamalarını anlamada kardiyologlara ve sağlık uzmanlarına rehberlik etmeyi amaçlamaktadır.

6)Çalışmanın Dizaynı:
ChatGPT dört defa farklı şekilde yönlendirilmiş ve ardından Amerikan Kalp Derneği'nin (AHA) AF ile ilgili sık sorulan sorularından türetilen 16 özdeş soru sorulmuştur2. Tüm yanıtlar için ChatGPT sürüm 3.5 kullanılmıştır. Spesifik istemler ve uyarımlar Tablo 1'de ayrıntılı olarak verilmektedir. Bu spesifik istemler; ChatGPT’yi aynı sorulan sorulara farklı bilgi, birikim ve kanıt düzeyinde cevap vermesine olanak tanımak için ön şartlandırmayı amaçlamaktadır. Figür 1’de her bir farklı istem için sorulan 16 adet soru ChatGPT'den alınan yanıtlar gözden geçirilerek yanlış, kısmen doğru, doğru veya mükemmel (referanslarla birlikte doğru olarak tanımlanmaktadır) olarak puanlanmıştır. Yanıtlar, AHA yanıtlarındaki bilgilerin %50'sinden daha azını içeriyorsa veya herhangi bir bilgi yanlış ise yanlış olarak işaretlenmiştir. Yanıtlar AHA'daki bilgilerin %50 ile %99'unu içeriyorsa ve yanlış bilgi içermiyorsa bu yanıtlar kısmen doğru olarak işaretlenmiştir. ChatGPT'den alınan bilgiler AHA yanıtlarındaki tüm bilgileri içeriyorsa ve herhangi bir ekstra bilgi doğru ise yanıtlar doğru olarak işaretlenmiştir. Son olarak yanıtlar doğru yanıt kriterlerini karşılıyor ve aynı zamanda istenen referansları ve/veya istatistikleri içeriyorsa bu yanıtlar mükemmel olarak işaretlenmiştir.

Tablo 1: Her form için kullanılan istemler

Form numarası	Form adı	ChatGPT istemi sağlanmak için verilen komutlar
1	Yönlendirme yok	Yönlendirme yok
2	Hasta dostu yönlendirme	Ben atriyal fibrilasyon hakkında daha fazla bilgi edinmeye çalışan bir hastayım. Size atriyal fibrilasyonla ilgili 16 soru soracağım. Lütfen benim anlamam için uygun olan dili kullanın, ancak yanıtlarınızın doğruluğundan ödün vermeyin. Cevaplarınızda mümkün olduğunca spesifik olun.
3	Hekim düzeyinde yönlendirme	Ben atriyal fibrilasyon hakkında en güncel bilgileri öğrenmeye çalışan sertifikalı bir hekimim. Size atriyal fibrilasyon ile ilgili 16 soru soracağım. Lütfen tıbbi kavramları uzman düzeyinde anlamam için uygun olan dili kullanın. Cevaplarınızda mümkün olduğunca spesifik olun.
4	İstatistik istemi ve referanslar	Size atriyal fibrilasyon ile ilgili 16 soru soracağım. Vereceğiniz her cevap için ilgili istatistik, sayı veya hesaplamaları eklediğinizden emin olun. Cevaplarınız yayınlanmış tıbbi literatürden gelmeli ve cevaplarınızda bunlara atıfta bulunmalısınız.

Figür 1: Her bir istem için sorulan 16 soru



7) Çalışmanın Sonuçları:
Tüm formlarda, puanlama frekansları bir (%1,6) yanlış, beş (%7,8) kısmen doğru, 55 (%85,9) doğru ve üç (%4,7) mükemmel şeklinde sonuçlanmıştır. Analiz sonucunda, en az doğru (yani, doğru veya mükemmel) olarak kategorize edilen yanıtların oranlarının istem formuna göre önemli ölçüde farklılık göstermediğini ortaya çıkmıştır ($p = 0.350$). Ancak, Form 4 önemli ölçüde daha fazla mükemmel yanıta sahip olarak bulunmuştur ($p < 0.023$). Ortalama not düzeyi Form 1 için 14.23 ± 2.34, Form 2 için 12.81 ± 3.38, Form 3 için 16.73 ± 2.65 ve Form 4 için 14.85 ± 2.76'dır. Hasta dostu yönlendirme olan Form 2 yanıtları, Form 1, 3 ve 4'e kıyasla daha düşük bir ortalama not seviyesi (12.81 ± 3.38) göstermiştir ($p < 0.05$). Ortalama kelime sayısı Form 1 için 69,50 ± 22,72, Form 2 için 59,94 ± 12,49, Form 3 için 73,50 ± 9,27 ve Form 4 için 118,80 ± 36,41'dir. İstatistik/referans istemi olan Form 4 yanıtları önemli ölçüde daha uzun olarak saptandı ($p < 0.0001$). Tüm formlarda, referanslar üç yanıtta (%4,7) verilmiştir. Kaynak veya referans istendiğinde ChatGPT'nin 16 yanıttan sadece

üçünde (%18,8) kaynak belirtmesi dikkat çekicidir. ChatGPT tarafından toplam üç kaynak verilmiştir; iki kaynak akademik web sitelerinden ve bir kaynak da yayınlanmış literatürden alınmıştır.

8) Çalışma Hakkında Yorumlar:
Genel olarak ChatGPT, yönlendirmeden bağımsız olarak AF ile ilgili çoğu soruya doğru ve kapsamlı yanıtlar vermiştir. ChatGPT, yanıtların %90,6'sına en az "doğru" yanıt verirken, yanıtların %98,4'üne en az "kısmen doğru" yanıt vermiştir. ChatGPT ve tüm yapay zeka sohbet robotlarının kullanımındaki artış göz önüne alındığında, genel doğruluk hasta eğitimi için olumlu olarak görülebilir. Hastaların genel populasyonu geniş bir sağlık okuryazarlığı yelpazesine sahiptir. Belirli bir kişi muhtemelen sağlık okuryazarlığını doğrudan belirtmeyecek olsa da, örneğin bu çalışmada tanımlandığı gibi "sınıf seviyesi", bir konuşma bağlamında hastaların daha yüksek veya daha düşük bir sağlık okuryazarlığı seviyesi gösterebileceği öngörülebilir.

Ulusal Sağlık Enstitüleri (NIH) ve Amerikan Tıp Birliği (AMA) hasta eğitim materyallerinin altıncı ve sekizinci sınıf okuma seviyeleri arasında yazılmasını tavsiye etmektedir. 3 Çalışmamız ChatGPT yanıtlarının bu sınıf düzeyi tavsiyesini aştığını ve ortalama 14,66 FK puanı ile üniversite okuma düzeyinde olduğunu ortaya koymuştur. ChatGPT'nin sağlıkla ilgili sorulara yanıt verirken sınıf seviyesini değerlendiren önceki çalışmalarda da benzer sonuçlar bulunmuştur.4

Bu çalışmadaki veriler, ChatGPT'nin konuşmanın istemine bağlı olarak yanıtlarının karmaşıklığını değiştirdiğini desteklemektedir. Hasta olarak yönlendirildiğinde ChatGPT, hekim ya da araştırmacı olarak yönlendirilmediğinde ya da yönlendirildiğinde daha düşük sınıf seviyesinde yanıtlar vermektedir. Bu sonuçlar hala önerilen altıncı veya sekizinci sınıf okuma seviyesi tavsiyesini karşılamasa da, ChatGPT'nin yanıt karmaşıklığını değiştirebilmesi açısından oldukça önemlidir. Önceki çalışmalar ChatGPT'nin farklı karmaşıklıklardaki tıbbi bilgileri anlama yeteneğini göstermiştir; ancak, bir yapay zeka sohbet robotunun girdiye dayalı olarak karmaşıklığını değiştirme yeteneği biraz tanımsız ya da belirsiz kalmaktadır. Yanıtların %1,6'sında ChatGPT hastalara yanlış bilgi sunmuştur. YZ sohbet robotlarının bir dezavantajı, "YZ halüsinasyonları" olarak bilinen yanlış bilgileri güvenle sunma eğilimidir.5 Bu durum, kullanıcıların, özellikle de hastaların verilen bilgileri eleştirel bir gözle değerlendirmeleri ve teyit için sağlık uzmanlarına danışmaları gerektiğinin altını çizmektedir.

Bu çalışma ChatGPT'nin doğru ve uyarlanabilir bir tıbbi bilgi kaynağı sağladığını göstermektedir. Bununla birlikte, ChatGPT çalışmasının kendine özgü sınırlamaları vardır. Hastalar tıbbi soruları sonsuz sayıda şekilde girebilir ve bu da bu çalışmada görülmeyen yanıtlar üretebilir. Ayrıca, yapay zeka dil botları yanıtlarını genellikle önceki girdilere göre değiştirir, bu da bu çalışmada görülen doğruluk ve anlama düzeyini değiştirebilir. Daha geniş ölçekli birçok klinik araştırmaya ihtiyaç duyulmaktadır ve bu araştırmalar daha geniş bir soru yelpazesini içermeli ve farklı YZ modellerinden alınan yanıtları karşılaştırmalıdır.

Sonuç olarak; Bu çalışma, ChatGPT'nin hasta eğitimi için tamamlayıcı bir kaynak olarak önemli bir potansiyele sahip olduğunu göstermektedir. Sohbet robotu, nadiren yanlış bilgilendirme yapmış olsa da yüksek derecede doğrulukla yanıt vermesi oldukça önemlidir. ChatGPT'nin yanıt karmaşıklığını konuşma bağlamına göre uyarlama yeteneği, farklı sağlık okuryazarlığı düzeylerine sahip bireylere hitap etmesini sağlayan dikkate değer bir güçtür. Bu yüksek güvenilirlik ve okuyucu uyarlanabilirliği göz önüne alındığında, sağlık hizmeti sağlayıcılarının ChatGPT'yi bir bilgilendirme kaynağı olarak tavsiye etmeleri ancak yanlış verilerle karşılaşmanın küçük ama olası riskini de kabul etmeleri makul bir sonuç olacaktır.

Kaynaklar

1. Lee T J, Campbell D J, Rao A K, et al. (June 04, 2024) Evaluating ChatGPT Responses on Atrial Fibrillation for Patient Education. Cureus 16(6): e61680. DOI 10.7759/cureus.61680
2. American Heart Association: FAQ about AFib. (2024). Accessed: March 3, 2024: http://www.heart.org/en/health-topics/atrial-fibrillation.
3. Rooney MK, Santiago G, Perni S, et al.: Readability of patient education materials from high-impact medicaljournals: a 20-year analysis. J Patient Exp. 2021, 8:2374373521998847. 10.1177/2374373521998847
4. Lee TJ, Campbell DJ, Patel S, Hossain A, Radfar N, Siddiqui E, Gardin JM: Unlocking health literacy: the ultimate guide to hypertension education from ChatGPT versus Google Gemini. Cureus. 2024, 16:e59898.10.7759/cureus.59898
5. Hatem R, Simmons B, Thornton JE: A call to address AI "hallucinations" and how healthcare professionals can mitigate their risks. Cureus. 2023, 15:e44720. 10.7759/cureus.44720